APB-UC3M at eRisk 2024: Natural Language Processing and Deep Learning for the Early Detection of Mental Disorders

Notebook for the eRisk Lab at CLEF 2024

Alejandro Pardo Bacuñana[1,*], Isabel Segura Bedmar[1]

[1]Universidad Carlos III de Madrid (UC3M), Avenida de la Universidad 30, 28911, Leganés, Madrid, Spain

Abstract

This paper presents our participation in the CLEF eRisk 2024 competition, where we focus on the early detection of anorexia, eating disorders, and depression from social media data. For the first task (search for symptoms of depression), we explore different sentence semantic similarity models, achieving robust performance in identifying early depressive symptoms, achieving the second best results in most of the evaluation metrics.

For the second task (early detection of signs of anorexia), we use an ensemble of traditional machine learning algorithms. In the eating disorders detection task, we use contextualized embeddings from BERT to represent the texts, and then, classify them with a neural network.

The findings highlight the posibility and room for improvement of early intervention and the potential of social media analysis to provide timely support for individuals at risk.

Keywords

Transformers, Deep Learning, Neuronal Networks, Mental Disorders, Early Detection, eRisk, CEUR-WS

1. Introduction

The main goal of the CLEF eRisk 2024 competition is to promote the development of advanced approaches for the early detection of various mental health issues through social media analysis [1, 2].

This competition, part of the Conference and Labs of the Evaluation Forum (CLEF), has been a significant event since its inception in 2017. It brings together researchers and practitioners to collaborate on innovative solutions for identifying early signs of mental health disorders from social media data. By analyzing textual content from social media posts, we can gain valuable insights into individuals' mental states and potentially provide early warnings for those at risk. Early detection is crucial for various applications, from identifying potential sexual offenders to detecting victims of suicidal tendencies, enabling interventions before it is too late [3].

Our participation in the CLEF eRisk 2024 competition focuses on the development and refinement of models to detect early signs of anorexia, eating disorders, and depression. Early intervention in these cases is crucial for providing timely support and improving outcomes for affected individuals [1, 2]. These conditions, prevalent among various demographics, often manifest in subtle linguistic cues that can be identified through sophisticated text analysis techniques.

In the last decade, social media has become a vital platform for individuals to express their thoughts, emotions, and ideas. This has opened up new avenues for the analysis of online data, which can be leveraged for numerous purposes such as business and marketing strategies, political planning, stock market predictions, and emergency awareness [4, 5].

In the healthcare domain, social media posts have been instrumental in detecting disease outbreaks, identifying smoking patterns, and recognizing adverse drug reactions, among others [6, 7]. More recently, the automatic detection of mental health issues has gained considerable attention within

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

alejandro.p.bascunana@alumnos.uc3m.es (A. Pardo Bacuñana); isegura@inf.uc3m.es (I. Segura Bedmar)

0009-0000-9749-637X (A. Pardo Bacuñana); 0000-0002-7810-2360 (I. Segura Bedmar)

the field of Natural Language Processing (NLP) [8]. Platforms like Twitter, Facebook, blogs, online forums, and Reddit provide rich corpora for detecting various mental health problems, including anxiety, depression, suicidal thoughts, and eating disorders [9].

In this paper, we present our approaches and methodologies for each of the three tasks of eRisk 2024: search for symptoms of depression (task 1), early detection of signs of anorexia (task 2), and measuring the severity of the signs of eating disorders (task 3). We employ a variety of NLP techniques like BERT for creating word embeddings, Machine Learning, and Deep Learning, to develop robust models capable of accurately detecting signs of anorexia, eating disorders, and depression. For task 1, we we achieved high accuracy in the early detection of anorexia.

2. Approaches and Experiments for each of the Tasks

In this section, we describe the different approaches that we have used for each one of the tasks.

2.1. Methods for task 1: search for symptoms of depression

The first task involved ranking and classifying sentences from a collection of user posts based on their relevance to symptoms of depression listed in the Beck Depression Inventory-II (BDI-II) questionnaire [10]. Participants had to provide rankings for all 21 depression symptoms in the BDI-II. A sentence was considered relevant to a depression symptom when it conveyed information about the user's condition or state related to that symptom. In other words, a sentence could be relevant even if it indicated that the user did not exhibit that particular symptom.

We explored three different approaches: the first one based on sentence semantic similarity models, the second one based on a RoBERTa classifier model and an ensemble that combined the previous ones.

2.1.1. Sentence Semantic Similarity Models

Our semantic similarity approach is based on Sentence Transformers [11], which leverage transformer models [12], specifically BERT [13]. These architectures are able, on the one hand, to capture the semantic relationships between words and tags; and on the other hand, to handle ambiguity in the text, as they considers the context of the words to represent them.

Thus, these models allow us to obtain representations of the sentences (embeddings) so that later, by using mathematical formulas such as cosine similarity, the level of similarity of meanings between both texts can be extracted.

Cosine similarity is a metric used to determine the similarity between two vectors. It is calculated as the cosine of the angle between the first and second vectors. A value of 1 indicates that the vectors are identical, 0 means that they are orthogonal (unrelated), and -1 implies that they are opposite. It is useful in semantic textual similarity, semantic search, or paraphrasing [14, 15].

In this approach, the following Sentence Transformers models were used:

- **all-MPNet-base-v2:** To obtain this model, the *microsoft/mpnet-base* model was fine-tuned on a dataset of 1B sentence pairs using a self-supervised contrasting learning objective. This particular model maps sentences and paragraphs to a dense vector space of 768 dimensions and is typically used for tasks such as clustering or semantic search[16].
- **all-MiniLM-L12-v2:** This model maps sentences and paragraphs to a dense 384-dimensional vector space and can be used for tasks such as clustering or semantic search. It is also derived from the pre-trained*microsoft/MiniML-L12-H384-uncased* model [17], which was fine-tuned using a dataset of 1B sentence pairs.
- **all-MiniLM-L6-v2:** a similar model to the previous one, the main difference being the number of hidden layers, 6 instead of 12. It has as its foundation the pre-trained model *nreimers/MiniLM-L6-H384-uncased* which in turn was again based on the Microsoft model *MiniLM-L12-H384-uncased*. [17].

Each model was used to calculate the numerical representations [18] of each sentence to be classified. Thus, its label is determined according to the highest similarity index to the annotated sentences in the training set. The sentences used for calculating the similarity were all the ones in the training set plus the eligible answers to each question of the BDI-II questionnaire [10].

In other words, if a sentence has a high similarity to the sentences indicating sadness (symptom 1), it is classified as such. For example, suppose we have the sentence "I am very sad, really, sad sad sad" which belongs to symptom 1 ("Sadness"), the model would calculate its numerical representation and calculate the cosine similarity with those of the reference sentences indicating sadness (label 1). If the similarity is high with sentences labelled under the same symptom (1), the sentence is classified as indicative of sadness with label "1".

Apart from the classification, the model also returned a decimal number between 1 and 10, reinterpreted from its cosine similarity (since cosine similarity returns a decimal number from 0 to 1), as the degree of relevance for that symptom in the BDI-II [10] questionnaire. This means that for example, for symptom 1 ("Sadness") the sentence "I feel sad all the time" would return 8.5, while the sentence "I did not feel sad in a long time" would return 2.7.

It is important to clarify that these models have a disadvantage in that they only label sentences once, i.e. sentence that may be related with more than one symptom at a time are not classified again, these models only address a multi-classification task but not multi-labelling. For example for the sentence "I am sad and crying" relevant for symptom 1 ("Sadness") and 10 ("Crying"), it would be classified as "1" or "10" with a single degree of relevance.

2.1.2. Classifier model RoBERTa

In our second approach, we deal with the first task as a multi-labelling classification problem.

The architecture chosen was RoBERTa (a Robustly Optimised BERT Approach) [19], which is a variant of the BERT model. Specifically, for this approach, the pre-trained SamLowe/roberta-base model was used, which is based on the previously described architecture. This model was designed for a multi-label classification task in sentiment analysis. The dataset chosen to train the model was *goemotions* [20]. The corpus is a set of multi-labeled [21] texts based on postings on the social network Reddit, where one or more labels can be applied to any given input text, such labels being a different type of both negative and positive emotion from the following list: "Disappointment, sadness, annoyance, neutral, disapproval, realisation, nervousness, approval, joy, anger, embarrassment, caring, remorse, disgust, grief, confusion, confusion, relief, desire, admiration, optimism, fear, love, excitement, curiosity, amusement, surprise, gratitude, pride".

In our study, we adjusted the model to predict only 21 labels (instead of the 28 in the original model), which are the 21 symptoms described in the BDI-II [10] test present in the symptom table. The model was then fine-tuned on the training dataset of the task.

The fitted model in addition to the classification, also provides the probability of each predicted label. This probability provided us with a relevance level for the test symptom in which the sentence was classified.

2.1.3. Ensemble

We also explored the combination of the previous approaches to deal with the task. We studied different ensembles of the sentence semantic models and the classifier RoBERTa. Based on our results during the development phase, our final submission was formed by aggregating the results from the *all-MPNet-base-v2*, *all-MiniLM-L12-v2*, and RoBERTa classifier models by majority voting and averaging the ranking value.

2.2. Methods for task 2: early detection of signs of anorexia

For this task, the goal was to perform binary classification of users, determining whether they were at risk for anorexia or not. The model had to analyze the user's sentences sequentially. If at any point the

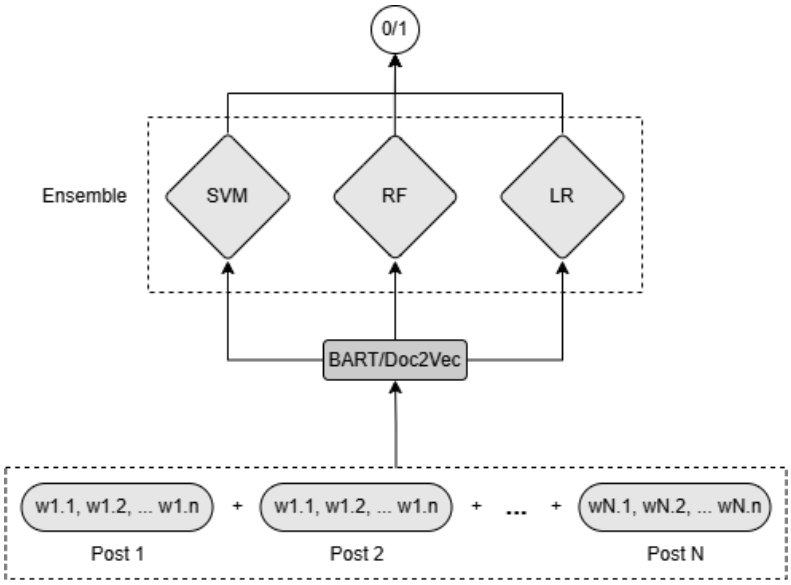

Figure 1: Architecture of the model for the task 2

model deemed the user to be at risk for anorexia based on the sentences, it had to issue an alert. This alert was communicated to a server. Importantly, once a positive risk alert was submitted for a user, it could not be changed or retracted in later stages of analyzing that user's data.

This binary classification task was tackled using a ensemble model of three classical algorithms, in addition to experimenting with different text representations. The ensemble models combine the predictions of several models to produce a final prediction, which can help reduce overfitting and improve generalisation by averaging the weaknesses and strengths of each individual model.

Specifically, our ensemble is composed of three different models: Support Vector Machines (SVM), Logistic Regression (LR) and Random Forest (RF). The texts were represented using embeddings produced by BART [22] and Doc2Vec [23].

Doc2Vec is an extension of the Word2Vec [24] model that allows the representation of documents as fixed-length vectors in a high-dimensional space. To do this, a neural network [25] is trained on a large corpus of text, where the network learns to predict words based on the surrounding context. As a result, documents with similar content or context will have similar vector representations, making it easier to identify relationships and patterns.

BART is a transformer model [12] that is pre-trained as a denoiser autoencoder, which means that its pre-training consisted of two stages. A first one where the text was corrupted with an arbitrary noise function and a second one where a sequence-sequence model is subsequently learned to reconstruct the original text. Thus allowing it to learn a rich and contextual representation of the input data.

Its architecture is characterised by an encoder with an approach similar to the BERT (Bidirectional Encoder Representation from Transformers) [13] and a decoder that follows the style of the Generative Pre-trained Transformer (GPT) model [26]. Both base models and the transformer technology [12].

The choice of the best model was made on the basis of a GridSearch with a 10-split StratifiedKFold choosing its F1 as the best metric. The hyperparameters of the best models composing the ensemble with Doc2Vec-based text representation were as follows:

- Doc2Vec: "vector_size": 100, "window": 5, "workers": 4, "min_count":2, "epochs": 40.
- SVM: "C": 1,4, "kernel": "rbf".
- LR: "C": 0.0012, "class_weight": "balanced", "penalty": "l2", "solver": "liblinear".
- Random Forest: "max_depth": 10, "min_samples_split": 5, "n_estimators": 120.

On the other hand, the hyperparameters chosen in the ensemble with text representation based on BART were as follows:

- SVM: “C”: 1,8, “degree”: 4, “kernel”: “poly”.
- LR: “C”: 0.0012, “class_weight”: “balanced”, “penalty”: “l2”, “solver”: “liblinear”.
- Random Forest: “max_depth”: 11, “min_samples_split”: 2, “n_estimators”: 140.

Two submissions were made for this tasks, being run 0 the one for the model with text representation based on Doc2Vec and run 1 for the one with BART embeddings.

2.3. Methods for task 3: measuring the severity of the signs of eating disorders

The objective of this task was to estimate the levels of features associated with eating disorders based on user posts. This task is a continuation of the efforts made in 2022 and 2023. Participants were tasked with analyzing these postings to fill out a standard eating disorder questionnaire for each user. The questionnaire used in this task is the Eating Disorder Examination Questionnaire (EDE-Q) [27], specifically focusing on questions 1-12 and 19-28. The EDE-Q is a 28-item self-reported questionnaire adapted from the semi-structured Eating Disorder Examination (EDE) interview [28]. It is designed to assess the range and severity of features associated with eating disorders.

We utilized BERT (Bidirectional Encoder Representations from Transformers) to generate word embeddings for our text data, which were subsequently fed into a neural network for classification. Our neural network, was designed to process these embeddings and classify the text into one of 22 specified categories (possible answers to the eating disorder test). Its architecture consisted of an input layer taking the 768-dimensional BERT embeddings, followed by a fully connected layer mapping this to a 400-dimensional space, with a ReLU activation function. A dropout layer with a rate of 0.5 was used to prevent overfitting, followed by another fully connected layer reducing the dimensions to 200, again with ReLU activation. The final output layer mapped the 200-dimensional input to the 22 output classes. The model was optimized with cross-entropy loss and the Adam optimizer.

3. Results

The detailed explanation for the metrics utilized can be found at the erisk 2024 overview paper in [1, 2]. Along with the system descriptions, results and experiments of the other participating teams.

3.1. Results for task 1: search for symptoms of depression

In task number 1, there were 9 teams in total. Being the team named *NUS-IDS* the one that achieved the best results. Most of the teams performed at least 3 different runs (submissions).

Team	Run	AP	R-PREC	P@10	NDCG
APB-UC3M	APB-UC3M_all-MiniLM-L6-v2	0.354	0.391	0.986	0.591
APB-UC3M	APB-UC3M_all-MiniLM-L12-v2	0.337	0.378	**0.990**	0.564
APB-UC3M	APB-UC3M_sentsim-all-mpnet-base-v2	0.293	0.330	0.967	0.525
APB-UC3M	APB-UC3M_ensemble	0.057	0.120	0.324	0.191
APB-UC3M	APB-UC3M_classifier_roberta-base-go_emotions	0.056	0.118	0.371	0.206
NUS-IDS	Config 5	**0.375**	**0.434**	0.924	**0.631**

Table 1
Ranking-based evaluation for Task 1 (majority voting). In bold, best results for each metric.

As can be seen in Tables 1 and 2, the semantic similarity models have been the ones that have obtained the best metrics, far above the classifier model RoBERTa.

Both unanimously and by majority, the semantic similarity models have achieved the second best metrics in Average Precision, R-PREC, NDCG and P@10. Even the model *all-MiniLM-L12-v2* obtained the best metric of P@10 in majority labelling.

This indicates that semantic similarity models have been able to generalise much better than the classifier model RoBERTa with texts never seen before. In addition, a better metric on the precision in

Team	Run	AP	R-PREC	P@10	NDCG
APB-UC3M	APB-UC3M_all-MiniLM-L6-v2	0.345	0.407	0.829	0.630
APB-UC3M	APB-UC3M_all-MiniLM-L12-v2	0.333	0.389	0.805	0.608
APB-UC3M	APB-UC3M_sentsim-all-mpnet-base-v2	0.285	0.342	0.776	0.561
APB-UC3M	APB-UC3M_ensemble	0.052	0.106	0.248	0.193
APB-UC3M	APB-UC3M_classifier_roberta-base-go_emotions	0.033	0.084	0.190	0.169
NUS-IDS	Config 5	**0.392**	**0.436**	0.795	**0.692**
MeVer-REBECCA	CosineSimilarity gpt	0.305	0.357	**0.833**	0.551

Table 2
Ranking-based evaluation for Task 1 (unanimity). In bold, best results for each metric.

10 (precision on the first 10 ranked items of each symptom) indicates that the method used to calculate the ranking of the sentence classification has been quite successful.

As for the classifier model, it has been shown that it has not been able to generalise as well as it should. A solution to this problem could be to train it for a longer time with a larger collection of sentences in order to try to improve its generalisation.

Finally, the ensemble model did not perform as expected. Its score was much lower than that of the models it incorporated. One of the reasons could be the decision to eliminate the multi-labelled sentences, we only kept the first occurrence of the sentence and its label.

3.2. Results for task 2: early detection of signs of anorexia

In task number 2, there were 10 teams in total. Being the team named *NLP-UNED* the one that achieved the best results. Most of the teams performed 5 different runs (submissions).

team	#runs	#user writings processed (from 1st to last response)	lapse of time
BioNLP-IISERB	5	10	09:39
GVIS	5	352	3 days 12:36
Riewe-Perla	5	2001	2 days 11:25
UNSL	3	2001	07:00
UMUTeam	5	2001	06:34
COS-470-Team-2	5	1	-
ELiRF-UPV	4	2001	12:27
NLP-UNED	5	2001	09:40
SINAI	5	2001	3 days 23:49
APB-UC3M	2	2001	6 days 21:34

Table 3
Task 2 (anorexia): participating teams, number of runs, number of user writings processed by the team, and lapse of time taken for the entire process.

Table 3 shows the time efficiency data for each of the models of each team after processing the total number of available posts (2,001). One reason for the longer elapsed time compared to other systems is the inability of the Doc2Vec-based text rendering model to use *NVIDIA CUDA* (Compute Unified Device Architecture) [29]. CUDA is a parallel computing platform and application programming interface model created by NVIDIA that allows developers to use NVIDIA GPUs (graphics processing units) for general-purpose processing such as training artificial intelligence models.

In Table 4, we can observe the results obtained for the decision-based evaluation (how accurate were the classifications into anorexics and non-anorexics) and the ranking-based evaluation (how accurate was the numerical value associated with each post based on its relevance). The model with text representation based on *Doc2Vec* although far behind the other teams, was able to rank the sentences more accurately than the text model based on BART [22]. However, it is also observed that the model performs worse on the speed-based metrics (*latencyTP* and *speed*).

Team	Run	P	R	F1	$ERDE_5$	$ERDE_{50}$	latencyTP	speed	latency-weighted F1
APB-UC3M	0	0.17	0.99	0.28	0.15	0.08	9.00	0.97	0.28
APB-UC3M	1	0.15	0.99	0.26	0.13	0.09	2.00	1.00	0.26
NLP-UNED	1	0.67	0.97	**0.79**	0.09	0.04	14.00	0.95	**0.75**
BioNLP-IISERB	4	**0.73**	0.62	0.67	0.08	0.05	4.00	0.99	0.66
Riewe-Perla	0	0.45	0.97	0.62	**0.07**	**0.02**	6.00	0.98	0.60

Table 4
Decision-based evaluation for Task 2. In bold, best results for each metric.

		1 writing			100 writings			500 writings			1000 writings		
Team	Run	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100
APB-UC3M	0	0.00	0.00	0.03	0.40	0.56	0.26	0.00	0.00	0.09	0.00	0.00	0.13
APB-UC3M	1	0.10	0.06	0.07	0.00	0.00	0.18	0.00	0.00	0.10	0.00	0.00	0.08
UNSL	1	**1.00**	**1.00**	**0.69**	**1.00**	**1.00**	0.80	0.90	0.81	0.69	0.80	0.88	0.72
NLP-UNED	1	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.89	**1.00**	**1.00**	**0.92**	**1.00**	**1.00**	**0.92**
NLP-UNED	3	**1.00**	**1.00**	0.45	**1.00**	**1.00**	**0.91**	**1.00**	**1.00**	0.91	**1.00**	**1.00**	0.89

Table 5
Ranking-based evaluation for Task 2. In bold, best results for each metric.

There are several explanations why the presented models have not been able to obtain similar results to that of the other teams. In terms of speed, apart from the impossibility of using CUDA in the Doc2Vec-based model, there has also been a shortcoming in terms of the equipment used to communicate with the server. In other words, during the evaluation phase, there was no sufficiently powerful equipment available that could accelerate the models.

On the other hand, regarding the accuracy of the models, a possible solution could be to change the type of ensemble used so that it is not carried out by majority voting, but rather specific weights are applied to each of the models that make up the [30], thus giving more relevance to the models that obtained the best metrics individually (e.g.: SVM). It should also be noted that due to a failure in the local machine used to communicate with the server, some messages were lost that could have been key in making decisions.

3.3. Results for task 3: measuring the severity of the signs of eating disorders

In task number 3, there were 5 teams in total. Being the team named *SCaLAR-NITK* the one that achieved the best results. Most of the teams performed at least 2 different runs (submissions).

Results depicted in table 6 show our team, APB-UC3M, achieved a MAE of 2.003, which is slightly higher than some of the top-performing teams such as SCaLAR-NITK, whose best run had a MAE of 1.874. Our MZOE was 0.869, indicating that our model had a moderate number of exact prediction matches, while our MAE_{macro} was 2.142. Our GED score of 2.647 suggests that our model's predicted sequences were relatively close to the ground truth sequences.

In terms of subscale scores, our Restraint Score (RS) was 2.253, Eating Concern Score (ECS) was 1.884, Shape Concern Score (SCS) was 2.101, and Weight Concern Score (WCS) was 1.823. These results show that our model performed consistently across different subscales, although there is room for improvement, particularly when compared to the leading team's scores.

team	run ID	MAE	MZOE	MAE_{macro}	GED	RS	ECS	SCS	WCS
baseline	all 0s	3.790	0.813	4.254	4.472	3.869	4.479	4.363	3.361
baseline	all 6s	1.937	0.551	3.018	3.076	3.352	2.868	3.029	2.472
baseline	average	1.965	0.884	1.973	2.337	2.486	1.559	2.002	1.783
APB-UC3M	0	2.003	0.869	2.142	2.647	2.253	1.884	2.101	1.823
RELAI	0	2.331	0.914	2.243	2.394	2.222	2.324	2.340	**1.812**
SCaLAR-NITK	0	1.912	0.591	**1.643**	2.495	2.713	1.568	**1.536**	2.098
SCaLAR-NITK	1	1.980	0.664	1.972	2.570	2.562	**1.553**	1.960	2.066
SCaLAR-NITK	2	1.879	**0.568**	1.942	2.158	2.477	2.222	2.245	2.364
SCaLAR-NITK	3	1.932	0.586	1.868	**2.117**	2.430	2.046	2.242	2.407
SCaLAR-NITK	4	**1.874**	0.672	1.820	2.292	**2.140**	1.557	1.880	2.061

Table 6
Task 3 Results. Participating teams and runs with corresponding scores for the metrics. In bold, best results for each metric.

The baseline results, especially the "all 6s" run, had a surprisingly strong performance, with a MAE of 1.937, which is close to our own. This indicates that a simplistic approach can still achieve competitive results, underscoring the complexity of improving upon simple heuristics in this task.

Overall, our model demonstrated a consistent performance but did not achieve the top results. The reason behind it could be due to the simplicity of the neuronal network used for the prediction.

4. Conclusions and Future Work

We participated in all three tasks of the eRisk 2024 shared tasks [1, 2].

Our models demonstrated a reasonable ability to identify early signs of depression from social media posts (task 1). The performance metrics showed that while the models were effective to an extent, there is still a gap between our approach and the desired level of accuracy. The challenge lies in capturing the subtle and varied ways depression symptoms can manifest in online behavior. One future approach we think of could be the improvement of the data quality. Expanding the dataset to include a wider variety of social media platforms and types of user interactions by including data augmentation techniques [31] and other social media datasets would help create more comprehensive models capable of generalizing across different contexts and user behaviors.

The models developed for detecting anorexia achieved poor results (task 2), revealing the need for more experimentation in this specific topic. Similar to Task 1, the variability in how individuals express anorexic symptoms online poses a significant challenge. The nuanced language and diverse expressions of anorexia necessitate more sophisticated models that can understand context and subtext better. To improve our results across the task, we propose the incorporation of sequential models such as GRU (Gated Recurrent Units) [32] or LSTM (Long Short-Term Memory) [33] which could potentially enhance the performance of our models by better capturing temporal dependencies and the sequential nature of user posts.

For task 3, our results were promising, particularly when compared to baseline models. Our approach, which involved the use of BERT embeddings and a neural network, provided a deeper understanding of user behavior, yet there is room for enhancing the precision and recall of our predictions. Future improvements could focus on enhancing the model's ability to capture subtle nuances in user posts that are indicative of eating disorders, possibly through more advanced embedding techniques or more sophisticated neural network architectures.

Acknowledgments

This work was supported by ACCESS2MEET project (PID2020-116527RB-I0) supported by MCIN AEI/10.13039/501100011033/.

References

[1] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Lecture Notes in Computer Science, Springer, 2024.

[2] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum CLEF 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[3] M. C. Prince, L. Srinivas, A review and design of depression and suicide detection model through social media analytics, in: Proceedings of International Conference on Deep Learning, Computing and Intelligence: ICDCI 2021, Springer, 2022, pp. 443–455.

[4] M. Wankhade, A. C. S. Rao, C. Kulkarni, A survey on sentiment analysis methods, applications, and challenges, Artificial Intelligence Review 55 (2022) 5731–5780.

[5] S. Dorle, N. Pise, Political sentiment analysis through social media, in: 2018 second international conference on computing methodologies and communication (ICCMC), IEEE, 2018, pp. 869–873.

[6] M. Omar, D. Brin, B. Glicksberg, E. Klang, Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review, American Journal of Infection Control (2024).

[7] J.-Y. Lee, Y.-S. Lee, D. H. Kim, H. S. Lee, B. R. Yang, M. G. Kim, The use of social media in detecting drug safety–related new black box warnings, labeling changes, or withdrawals: scoping review, JMIR public health and surveillance 7 (2021) e30137.

[8] W. A. Gadzama, D. Gabi, M. S. Argungu, H. U. Suru, The use of machine learning and deep learning models in detecting depression on social media: A systematic literature review, Personalized Medicine in Psychiatry 45 (2024) 100125.

[9] A. Ahmed, S. Aziz, C. T. Toro, M. Alzubaidi, S. Irshaidat, H. A. Serhan, A. A. Abd-Alrazaq, M. Househ, Machine learning models to detect anxiety and depression through social media: A scoping review, Computer Methods and Programs in Biomedicine Update 2 (2022) 100066.

[10] A. T. Beck, R. A. Steer, G. Brown, Beck depression inventory–ii, Psychological assessment (1996). URL: https://doi.org/10.1037/t00742-000.

[11] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, CoRR abs/1908.10084 (2019). URL: http://arxiv.org/abs/1908.10084. arXiv:1908.10084.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[14] L. Muflikhah, B. Baharudin, Document clustering using concept space and cosine similarity measurement, in: 2009 International conference on computer technology and development, volume 1, IEEE, 2009, pp. 58–62.

[15] B. Li, L. Han, Distance weighted cosine similarity measure for text classification, in: Intelligent Data Engineering and Automated Learning–IDEAL 2013: 14th International Conference, IDEAL 2013, Hefei, China, October 20-23, 2013. Proceedings 14, Springer, 2013, pp. 611–618.

[16] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, Advances in neural information processing systems 33 (2020) 16857–16867.

[17] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. arXiv:2002.10957.

[18] F. Almeida, G. Xexéo, Word embeddings: A survey, 2023. arXiv:1901.09069.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov,

Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019). URL: https://doi.org/10.48550/arXiv.1907.11692.

[20] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, Goemotions: A dataset of fine-grained emotions, 2020. arXiv:2005.00547.

[21] A. N. Tarekegn, M. Ullah, F. A. Cheikh, Deep learning for multi-label learning: A comprehensive survey, 2024. arXiv:2401.16549.

[22] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, CoRR abs/1910.13461 (2019). URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.

[23] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.

[24] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013. arXiv:1301.3781.

[25] J. Schmidhuber, Deep learning in neural networks: An overview, Neural networks 61 (2015) 85–117.

[26] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[27] R. Murphy, S. Straebler, Z. Cooper, C. G. Fairburn, Cognitive behavioral therapy for eating disorders, Psychiatric Clinics of North America 33 (2010) 611–627. URL: https://doi.org/10.1016/j.psc.2010.04.004, cognitive Behavioral Therapy.

[28] C. G. Fairburn, G. T. Wilson, K. Schleimer, Binge eating: Nature, assessment, and treatment, Guilford Press New York, 1993.

[29] R. S. Dehal, C. Munjal, A. A. Ansari, A. S. Kushwaha, Gpu computing revolution: Cuda, in: 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 197–201. doi:10.1109/ICACCCN.2018.8748495.

[30] I. D. Mienye, Y. Sun, A survey of ensemble learning: Concepts, algorithms, applications, and prospects, IEEE Access 10 (2022) 99129–99149. doi:10.1109/ACCESS.2022.3207287.

[31] L. Perez, J. Wang, The effectiveness of data augmentation in image classification using deep learning, CoRR abs/1712.04621 (2017). URL: http://arxiv.org/abs/1712.04621. arXiv:1712.04621.

[32] R. Dey, F. M. Salem, Gate-variants of gated recurrent unit (GRU) neural networks, CoRR abs/1701.05923 (2017). URL: http://arxiv.org/abs/1701.05923. arXiv:1701.05923.

[33] R. C. Staudemeyer, E. R. Morris, Understanding LSTM - a tutorial into long short-term memory recurrent neural networks, CoRR abs/1909.09586 (2019). URL: http://arxiv.org/abs/1909.09586. arXiv:1909.09586.